Nouvelles perspectives en sciences sociales

# Une approche textométrique pour étudier la transmission des savoirs biologiques au XIX$^{e}$ siècle

# A Textometrical Approach to Study the Transmission of Biological Knowledge in the XIX$^{th}$ Century

Philippe Gambette and Nadège Lechevrel

Volume 12, Number 1, November 2016

URI: https://id.erudit.org/iderudit/1038375ar
DOI: https://doi.org/10.7202/1038375ar

See table of contents

Publisher(s)

Prise de parole

ISSN

1712-8307 (print)
1918-7475 (digital)

Explore this journal

Cite this article

Gambette, P. & Lechevrel, N. (2016). Une approche textométrique pour étudier la transmission des savoirs biologiques au XIX$^{e}$ siècle. *Nouvelles perspectives en sciences sociales*, *12*(1), 221–253. https://doi.org/10.7202/1038375ar

Article abstract

This article addresses the issue of the relationship between qualitative and quantitative analysis (i.e. computerized analysis of textual data) through experiments (and their results) conducted in a research project focusing on the impact of biological knowledge on French literary work in the XIX$^{th}$ century (Biolographes: http://biolog.hypotheses.org). The first part sets out the practical aspects of digital corpora, from access to the texts, to their transformation by OCR, to the storage of their metadata. The second and third parts illustrate how textometrical tools and visualizations (TXM, TreeCloud) serve as a point d'appui to many working hypotheses. In conclusion, the article emphasizes the role played by NLP in computerized tools for literary analysis.

# Une approche textométrique pour étudier la transmission des savoirs biologiques au XIX$^{e}$ siècle[1]

**Philippe Gambette**
Université Paris-Est, LIGM (UMR 8049), UPEM, CNRS,
ESIEE, F-77454, Marne-la-Vallée

**Nadège Lechevrel**
Fondation Maison des sciences
de l'homme / ANR Biographes, Paris

## Introduction

Cet article présente des exemples d'application d'analyses textuelles informatisées avec TXM[2] et TreeCloud[3] réalisées

1 Ces recherches ont été conduites dans le cadre du projet *Biolographes*, projet ANR-13-FRAL-0013, 2014-2016, dirigé par Gisèle Séginger (http://biolog.hypotheses.org/).

2 TXM est un outil d'analyse quantitative et qualitative de corpus textuels numériques développé par une équipe de recherche de Lyon (http://textometrie.ens-lyon.fr/). Serge Heiden, Jean-Philippe Magué et Bénédicte Pincemin, « TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement », dans Sergio Bolasco, Isabella Chiari, Luca Giuliano (dir.), *Statistical Analysis of Textual Data, Proceedings of the 10th International Conference on Statistical Analysis of Textual Data (JADT 2010)*, Edizioni Universitarie di Lettere Economia Diritto, 2010, p. 1021-1032; Nadège Lechevrel, « Fouille de données textuelles et recherche documentaire automatiques pour l'histoire des théories linguistiques », dans Pascal Charbonnat, Mahé Ben Hamed, Guillaume Lecointre (dir.), *Apparenter la pensée ? Vers une phylogénie des concepts savants*, Matériologiques, 2014, p. 219-243.

3 TreeCloud est un outil qui permet de générer des nuages de mots arborés à partir d'un texte (http://www.treecloud.org). Philippe Gambette et Jean Véronis, « Visualising a Text with a Tree Cloud », dans Hermann Locarek-

dans le cadre du projet ANR « Biolographes : création littéraire et savoirs biologiques au XIX^e^ siècle », dirigé par Gisèle Séginger et Thomas Klinkert. L'équipe *Biolographes* a choisi de travailler sur des savoirs peu abordés alors qu'ils ont connu un succès croissant au XIX^e^ siècle, assurant la promotion d'une discipline qui se constitue progressivement, en se différenciant de l'histoire naturelle, la médecine et la physiologie. Le programme comporte à la fois un travail d'inventaire et l'établissement d'un corpus pour l'analyse des textes portant sur l'utilisation des savoirs biologiques dans la littérature française au XIX^e^ siècle : comment elle s'en empare, les utilise de manière plus ou moins fiable, les commente, les prolonge, etc[4]. Le corpus est provisoirement divisé en trois sous-corpus : un corpus principal composé d'œuvres littéraires et d'avant-textes, un premier corpus secondaire réunissant les œuvres scientifiques les plus marquantes du XIX^e^ siècle qui ont influencé les auteurs du corpus principal, ainsi qu'un second corpus secondaire rassemblant des articles de revues scientifiques et savantes et de vulgarisation[5]. Ce découpage est avant tout méthodologique, les chercheurs du projet entendant dépasser ces catégorisations (en genre littéraire / genre scientifique) pour souligner la circulation des savoirs et la co-construction des textes. Notre travail dans le cadre de ce projet ANR vise à explorer le corpus de textes à l'aide d'outils de textométrie et de visualisation. Nous donnons ici quelques résultats de ces

---

Junge et Claus Weihs (dir.), *Studies in Classification, Data Analysis, and Knowledge Organization, Proceedings of the International Federation of Classification Societies 2009 Conference (IFCS'09)*, n° 40, 2010, p. 561-570; Delphine Amstutz et Philippe Gambette, « Utilisation de la visualisation en nuage arboré pour l'analyse littéraire », dans Sergio Bolasco, Isabella Chiari et Luca Giuliano (dir.), *Statistical Analysis of Textual Data, Proceedings of the 10th International Conference on Statistical Analysis of Textual Data (JADT 2010)*, Edizioni Universitarie di Lettere Economia Diritto, 2010, p. 227-238.

4 Voir la présentation du programme sur le site du projet : http://biolog.hypotheses.org/.

5 Nadège Lechevrel, *Réception et vulgarisation des savoirs biologiques dans le corpus Biolographes*, 2015, billet de blog sur le carnet hypotheses.org du projet Biolographes, http://biolog.hypotheses.org/1276, site consulté le 25 mars 2016. Voir aussi http://biolog.hypotheses.org/category/corpus-biolographes.

premières explorations sur un sous-corpus de textes de réception et vulgarisation de travaux scientifiques.

## 1. Présentation des données

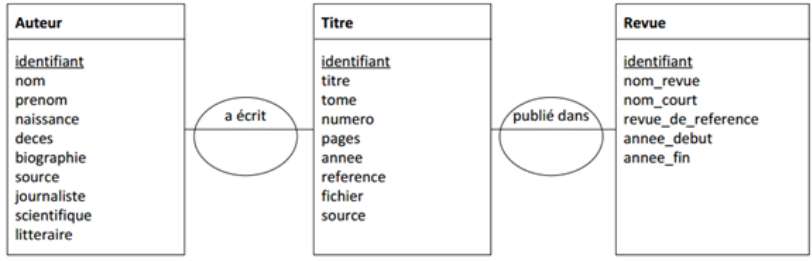
La diffusion des savoirs biologiques au XIX^e^ siècle assure la promotion d'une discipline qui se constitue progressivement en se différenciant de l'histoire naturelle, de la médecine et de la physiologie. Si les travaux sur l'histoire de la biologie ou sur l'histoire des idées (par exemple sur la circulation des métaphores de l'organisme aux XVIII^e^ et XIX^e^ siècles dans les sciences humaines; sur la réorganisation culturelle dans les sciences des XVIII^e^ et XIX^e^ siècles) sont nombreux, la critique littéraire n'a pas encore abordé de manière systématique et globale l'apport de la biologie naissante à la littérature française du XIX^e^ siècle. Pour étudier cette circulation des savoirs biologiques, les chercheurs du projet *Biolographes* se sont fixé pour objectif la constitution d'un corpus de textes de réception et/ou de vulgarisation par le dépouillement systématique d'une sélection de revues. Les textes de diffusion scientifique ont été prélevés dans des publications prestigieuses (telles la *Revue des deux mondes*, la *Philosophie positive*, etc.), des revues scientifiques, ou de vulgarisation.

La constitution d'un corpus entraîne automatiquement la constitution d'une base de données quelle qu'en soit la forme (tabulaire, relationnelle, à balises, etc.). Pour ce corpus de textes de réception, nous avons constitué une base de données sur trois niveaux (« Auteur », « Titre » et « Revue », voir la figure 1) qui permet de mettre en relation les auteurs (y compris leur pseudonyme) et leur biographie ainsi que leurs contributions dans différentes revues comme la *Revue des deux mondes* (1829-), la revue *Philosophie positive* (1867-1883), l'*Année biologique*, la *Science française* (1890-1901) ou la *Science illustrée* (1875-1877; 1888-1905). L'extraction de l'information pour la catégorie « Auteur » s'est faite via l'envoi de requêtes SPARQL de la Bibliothèque Nationale de France[6] et a permis de collecter des données bio-

[6] Utilisation du site http://data.bnf.fr/sparql. « Les données sont disponibles sur ce site, selon plusieurs syntaxes de RDF (Resource description framework),

graphiques telles que nom et prénom (normalisés), date de naissance et décès, et spécialités des savants. Les données « Titre » sont obtenues directement du dépouillement des revues réalisé par les membres de l'équipe : on place ensuite dans un tableau toutes les contributions par année et par auteurs selon les différentes revues. Enfin, les données « Revue » sont obtenues automatiquement par comptage : nombre d'articles par revue ainsi que périodisation, également en fonction des différentes appellations de la revue si cela s'applique (c'est le cas par exemple de la *Science illustrée* qui connaît trois périodes et deux noms différents, *La science illustrée* et *La science populaire*[7]).

Figure 1 : Modèle conceptuel de la base de données utilisée




Dans le cas où l'auteur n'est connu que par son nom de famille, une requête sans prénom donne des résultats à dépouiller manuellement. Quand une identité n'est pas trouvée, elle n'apparaît tout simplement pas dans les résultats. D'autres absences peuvent s'expliquer par des problèmes d'écriture du pseudonyme, traité par exemple dans la base de données de la Bibliothèque Nationale comme un seul bloc sous l'étiquette « nom de famille ».

soit RDF-XML, RDF-N3, et RDF-NT, ainsi qu'en JSON. Un document d'exemples de requêtes permet de se familiariser avec son interrogation » (http://data.bnf.fr/semanticweb). « La BnF a, depuis le 1er janvier 2014, placé ses métadonnées descriptives (données bibliographiques et d'autorité) sous la Licence Ouverte de l'État élaborée par la mission Etalab. L'utilisation de ces métadonnées est libre et gratuite sous réserve du maintien de la mention de leur source et de l'indication de leur date de récupération » (http://data.bnf.fr/licence).

7 Voir le billet correspondant sur http://biolog.hypotheses.org/1276.

Ainsi, le nom Dora d'Istria, balisé `<foaf:familyName>Dora d'Istria</foaf:familyName>`, avec les requêtes suivantes donne :

`{ ?auteur foaf:name «Dora d'Istria».}` : aucun résultat

`{ ?auteur foaf:familyName «Istria».}` : une liste de personnes dans laquelle ne figurera pas Dora D'Istria.

Pour certains auteurs de la revue la *Science française* ou la *Philosophie positive*, il est difficile de trouver des informations biographiques sur les auteurs et de savoir par exemple s'il s'agit de noms ou de pseudonymes. Il y a par ailleurs de nombreux textes aux auteurs anonymes dans notre corpus (6 sur 169 dans la *Revue des deux mondes*, 27 sur 120 dans la revue la *Science française*, 2 sur 6 pour *La nature*). Finalement, nous avons six pseudonymes célèbres (Alexandre Rameau, George Sand, Émile Saigey, Georges Béthuys, Diatomea pour Jules Girard et Justin d'Hennezis) et nous n'avons pas encore trouvé d'informations biographiques pour 41 auteurs en plus des auteurs anonymes, soit parce que nous ne possédons que leurs initiales (ex. V. K.), soit parce que le nom de famille est accompagné de l'initiale d'un prénom ne permettant pas d'identifier clairement l'auteur.

La base de données que nous venons de décrire permet plusieurs traitements, de la simple recherche documentaire à l'étude des réseaux littéraires/scientifiques. Des visualisations sur des plateformes déjà existantes, comme Palladio[8], une plateforme du projet *Networks in History*[9] de l'Université Stanford, permettront aux membres du projet d'obtenir à l'avenir un panorama visuel des auteurs présents dans notre base de données textuelles.

Les explorations qui suivent ont été réalisées sur le corpus des articles tirés de la *Revue des deux mondes* (1829-1971), dirigée par François Buloz (1803-1877) à partir de 1831, et connue pour être l'une des revues les plus lues par les savants et écrivains du XIX[e] siècle : entre 1857 et 1884, on y lit de nombreuses recensions

8 *Palladio, Humanities thinking with data*, http://palladio.designhumanities.org/, site consulté le 17 mai 2016.

9 *Networks in History: Data-driven tools for analyzing relationships across time*, http://hdlab.stanford.edu/projects/networks-in-history/, site consulté le 17 mai 2016.

de l'œuvre de Gustave Flaubert[10]; Paul de Rémusat (1831-1897), homme politique et grand vulgarisateur, y expose la théorie de la génération spontanée exposée par Félix Pouchet (1800-1872) dans son ouvrage *Hétérogénie* (1859); et Charles Martins (1806-1889), botaniste, y présente les travaux d'Ernst Haeckel (1834-1919) et Karl Vogt (1817-1895), donnant à découvrir le « monère » ou la vie des « polypes » si chers à Gustave Flaubert. Voici en outre (voir la figure 2) les domaines des « sciences » que la *Revue des deux mondes* couvrait :

Figure 2 : Liste des catégories « scientifiques » relative au classement des articles dans la *Revue des deux mondes*, produite dans la Table Générale établie en 1875 pour la période 1831-1874

| |
|---|
| sciences |
| médecine |
| sciences, médecine militaire |
| science, voyage scientifique |
| sciences, voyages |
| marine, sciences |
| sciences, agriculture |
| science hippique |
| sciences, agriculture, industrie de l'alimentation publique |
| sciences, philosophie |
| histoire des sciences |

Pour la période 1829-1901, 154 numéros de la *Revue des deux mondes* sont accessibles en mode texte sur Gallica qui a procédé à leur numérisation automatique. Cette numérisation est réalisée à l'aide de logiciels d'OCR qui traduisent des images de texte en texte. Dans ce contexte, on parle communément de taux de reconnaissance optique de caractères. Ce taux étant compris entre 60 et 99 % (selon Gallica), nous avons choisi de ne sélectionner dans notre corpus que des textes qui nécessitaient peu de

10 Philippe Dufour, « La feuille bulozienne », *Flaubert*, nº 9, 2013, http://flaubert.revues.org/2024, article consulté le 17 mai 2016.

correction[11], issus de Gallica, L'Observatoire de la vie littéraire[12] (*OBVIL*) ou Wikisource[13].

Notre corpus actuel (voir la figure 3) présente les caractéristiques suivantes :

Figure 3 : Description du corpus

Description des 168 textes du corpus *Revue des deux mondes* :
- 2 108 956 occurrences (taille des textes) et 54 030 formes uniques (volume du lexique) selon TXM
- 98 textes écrits par 19 auteurs « scientifiques »
- 41 textes écrits par 17 auteurs « gens de lettres »
- 29 textes écrits par des auteurs ayant les deux profils

Pour plus de détails sur le vocabulaire textométrique utilisé ici (taille, volume, mot, forme graphique, pivot, etc.), voir le glossaire du manuel en ligne de TXM : http://txm.sourceforge.net/doc/manual/manual84.xhtml.

Comme on le voit, les informations relatives aux spécialités des savants recueillies dans la base de données *data.bnf.fr* ont été exploitées pour partitionner les textes du corpus en deux types ou deux genres d'écriture, les textes de réception ou de vulgarisation écrits par des scientifiques (98 textes), d'une part, les textes écrits par des gens de lettres (écrivain, philosophe, historien, philologue, académicien, etc.) (41 textes), d'autre part. Les outils de textométrie s'accommodent d'une telle différence de taille de corpus[14], le nombre de mots des deux sous-corpus ayant le même ordre de grandeur. Une troisième catégorie, non utilisée dans cette partition, est réservée aux journalistes et vulgarisateurs de profession (nous nous sommes appuyés sur les informations extraites des données de la Bibliothèque Nationale de France

11 Nous avons pu constater deux types d'erreur liées à la qualité de l'image : d'une part, une résolution insuffisante qui génère des ambiguïtés sur certains caractères (virgule transformée en s, I transformés en 1, etc.), d'autre part, des altérations locales de l'image affectant une portion de texte nécessitant sa réécriture complète (marque d'un pli, défaut du matériel de numérisation, etc.).

12 OBVIL, *Observatoire de la vie littéraire*, http://obvil.paris-sorbonne.fr/corpus/critique/, site consulté le 17 mai 2016.

13 Wikisource, *Revue des deux mondes*, https://fr.wikisource.org/wiki/Revue_des_Deux_Mondes, site consulté le 17 mai 2016.

14 Pierre Lafon, « Sur la variabilité de la fréquence des formes dans un corpus », *Mots*, vol. 1, n° 1, p. 127-165, 1980.

pour cette catégorisation). L'étude contrastive des textes reposera ainsi principalement sur la catégorisation de leurs auteurs.

## 2. Aperçu des sujets d'intérêts scientifiques diffusés dans la *Revue des deux mondes*

Les revues savantes, scientifiques et de vulgarisation étaient lues et très faciles d'accès pour un public lettré et constituaient l'un des canaux principaux de diffusion des savoirs scientifiques auprès des écrivains. On fait donc l'hypothèse qu'en accordant une place aux textes de réception et de vulgarisation, il sera possible de montrer quels travaux scientifiques étaient les plus diffusés, participant ainsi à l'étude des modalités de transferts des savoirs biologiques vers les textes littéraires.

À partir du tableau de données « Titre », on obtient une première visualisation des thématiques abordées par les auteurs sélectionnés dans la *Revue des deux mondes*. Sur l'ensemble des titres, les formes graphiques « critique littéraire » et « revue scientifique » apparaissent de façon répétitive sans pour autant être informatives sur le contenu des articles qu'elles concernent, et sont donc retirées. Par ailleurs, l'application d'un antidictionnaire (c'est-à-dire une liste préétablie de 779 formes que l'on retire des textes à traiter, qui peut être adaptée en fonction du corpus étudié) fourni dans TreeCloud permet de ne garder que les formes pertinentes pour l'analyse des thématiques dans le résultat final. N'ayant pas détecté de vocabulaire spécifique du XIX$^{e}$ siècle parmi les mots les plus fréquents utilisés dans les analyses ci-dessous, cet antidictionnaire nous a semblé approprié. Par ailleurs, tous les mots sont passés en minuscules pour ne pas faire de différences avec ceux qui apparaissent en début de phrase. Enfin, bien qu'une lemmatisation soit possible dans TXM[15], qui

15 Cette lemmatisation peut toutefois conduire à des taux d'erreur supérieurs à ceux obtenus sur des textes en français contemporains. Par exemple, parmi les 65 formes du terme « éminens » dans le corpus, 9 sont reconnues comme noms alors que si l'on remplace ces formes par « éminents », elles sont toutes correctement reconnues comme adjectifs.

utilise TreeTagger[16] pour cela, nous avons choisi de travailler sur les formes graphiques suite à plusieurs exemples montrant la pertinence de travailler directement sur les formes graphiques (différence entre « science » et « sciences », « animal » et « animaux », etc.).

Les nuages arborés qui figurent ci-dessous, générés par le logiciel TreeCloud, regroupent un ensemble de mots en fonction de leur proximité dans le texte, selon la formule de cooccurrence Liddell[17]. La taille de ces mots reflète leur fréquence, ou bien l'indice de cooccurrence calculé dans TXM. Les arbres peuvent se lire indifféremment de droite à gauche et de gauche à droite. La longueur des branches est artificiellement fixée à une même valeur pour une meilleure lisibilité de l'arbre[18]. Les distances entre mots dans l'arbre ne peuvent donc pas être interprétées, et ce sont simplement les regroupements des mots au sein d'un même sous-arbre (par exemple, dans l'arbre de la figure 4, les mots « france », « lettres », « état » et « américain », qui correspondent à la série des « Lettres à un américain sur l'état des sciences en France ») que nous interprétons.

Enfin, dans les figures 4 et 5, les couleurs sont attribuées en fonction de la position moyenne de chaque mot : rouge vif (ou gris pâle dans cet article) pour le mot dont la position moyenne est la plus petite, bleu vif (ou noir dans cet article) pour celui dont la position moyenne est la plus grande. Ces couleurs indiquent donc la chronologie des mots dans les textes, le gris pâle reflétant les textes du début du XIX$^{e}$ siècle, le noir correspondant aux textes de la fin du XIX$^{e}$ siècle. Les mots peuvent aussi être colorés en fonction d'un score statistique selon un ensemble de 4 classes de couleur (rouge pour les scores les plus élevés, orange ensuite, puis bleu sombre, puis bleu pâle, qui

16 Helmut Schmid, « Probabilistic Part-of-Speech Tagging Using Decision Trees », *Proceedings of International Conference on New Methods in Language Processing*, 1994, p. 44-49.

17 Philippe Gambette, *User Manual for TreeCloud*, 2010, http://www.treecloud.org/DOWNLOADS/ManualTreecloud.pdf, p. 8.

18 Philippe Gambette, Nuria Gala et Alexis Nasr, « Longueur de branches et arbres de mots », *Corpus*, n° 11, 2012, p. 129-146.

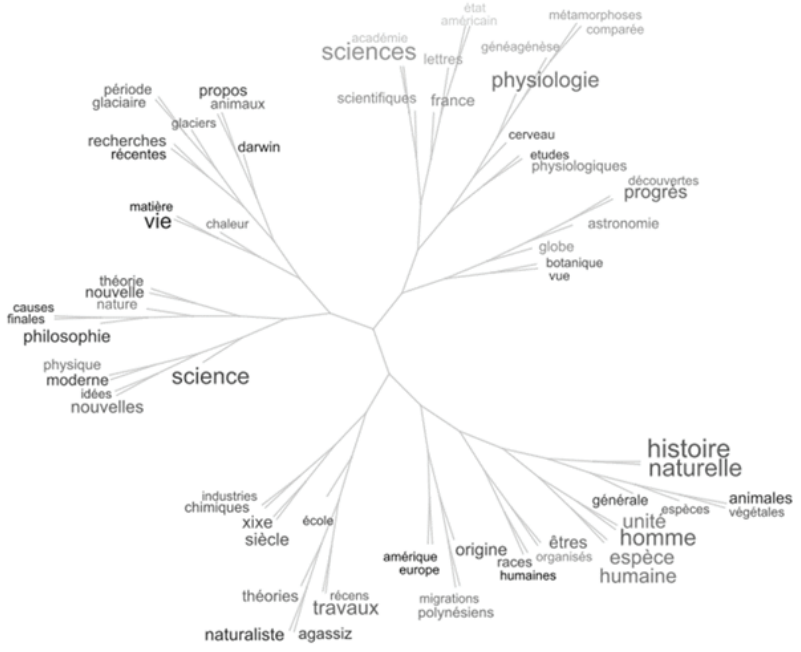
apparaissent dans cet article en niveaux de gris du plus foncé au plus pâle). Dans les figures 8 à 13, ce sont des indices de cooccurrence calculés par TXM qui sont utilisés : les arbres de ces figures ne contiennent que les cooccurrents d'un terme donné, c'est-à-dire les mots qui ont les plus hauts indices de cooccurrence avec ce terme. Ils traduisent une sur-représentation du mot dans le voisinage d'un mot cible.

Le nuage arboré des titres[19] (voir la figure 4) permet à un chercheur extérieur au projet *Biolographes* de se familiariser aussi bien avec les thématiques retenues dans le cadre du dépouillement de la *Revue des deux mondes* (« découvertes » et « progrès » scientifiques, « études physiologiques », évolutionnisme - « vie », « darwin », « science » et « philosophie », « histoire naturelle ») qu'avec les caractéristiques de la revue elle-même : la France et le reste du monde (« France », « Amérique », « Europe »), l'« Académie » et les « lettres » (ses lecteurs et contributeurs), et, de façon générale, un regard critique sur les travaux récents (« idées », « recherches », « théorie » d'un côté, « nouveau/-elle », « moderne », « récentes/récens » de l'autre).

Le nuage arboré montre en outre un découpage des savoirs biologiques où l'ancienne « histoire naturelle » agrège toujours, dans la *Revue des deux mondes*, la grande majorité des débats autour des espèces humaine, animale et végétale (« origine » et organisation - « organisés »), provoquant un éclatement de thématiques (plus ou moins unifiées en réalité) entre physiologie, recherches sur la vie (« darwin ») et travaux d'histoire naturelle. C'est en 1848, en France, que les contours de la discipline biologie se précisent, grâce en particulier à la Société de biologie de Paris fondée par Claude Bernard et Charles Robin (la revue *L'année biologique* est créée bien plus tard, en 1895, par Yves Delage (1854-1920)).

19 Les titres analysés ont une longueur de 10,7 mots en moyenne.

Figure 4 : Nuage arboré des mots présents trois fois ou plus (hors mots de l'antidictionnaire[20]) dans les titres des articles de la *Revue des deux mondes* (168 articles), colorés chronologiquement (gris clair pour les mots du début du siècle, noir pour les mots de la fin du siècle)




Le nuage arboré des 100 mots les plus fréquents de l'ensemble des textes du corpus (voir la figure 5) quant à lui permet de se rapprocher du contenu des articles, en ce qu'il représente davantage les interactions lexicales ou l'attirance contextuelle des mots. Si le premier nuage permettait de se faire une idée générale des sujets d'intérêt, celui-ci donne de l'information relative au thème, le rhème, en montrant comment les auteurs abordent ces grandes thématiques (toutes catégories confondues). S'il n'est pas possible de faire un lien direct entre tel sujet d'intérêt (voir la

20 Ces mots représentent 482 des 1477 occurrences des titres, alors que 736 occurrences de mots de l'antidictionnaire ont été ignorées : le nuage arboré correspond donc à près des deux tiers des occurrences du corpus de titres, hors antidictionnaire.

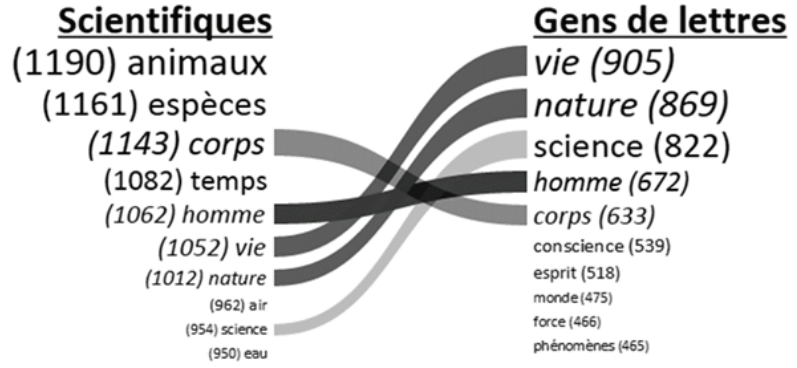
figure 6) et tel environnement lexical (voir la figure 7), les formes les plus fréquentes guident cependant l'analyse. On opérera en particulier trois regroupements : le premier marquant une forte présence de questionnements *philosophiques* et *métaphysiques* (l'inscription dans l'histoire autour du « temps » en haut à droite du nuage et le « doute » au centre à gauche de l'arbre), le deuxième portant sur *l'organisation* des espèces vivantes (l'« homme » et l'« animal »), le troisième relevant des *méthodes* et *principes* de LA science (« expérience », « faits », « raison » et « ordre » autour de la « science »; « conditions », « lois », « principe », « phénomènes » et « vie » autour de la « nature »). Enfin, les deux ensembles en bas, à gauche du nuage révèlent une approche classique du monde à travers l'étude de ses éléments (l'« eau », l'« air », la « terre », le « soleil ») pour l'un, et une focalisation sur les relations causales (« cause », « effet »), le mouvement et la force, des « parties », « organes », « élémens », « forme » et « corps » pour l'autre. (Voir la figure 5.)

Prenons enfin une liste de fréquences indiquant les 10 mots les plus fréquents dans les deux sous-corpus, les textes des scientifiques, d'une part, les textes des gens de lettres, d'autre part. À partir de cette liste, on repère rapidement les mots partagés par les deux sous-corpus. (Voir la figure 6.)

Figure 5 : Nuage arboré des 100 mots les plus fréquents[21] dans le corpus de textes de la *Revue des deux mondes*, colorés chronologiquement (gris pâle pour les mots du début du siècle, noir pour les mots de la fin du siècle)


21 Ces mots représentent 101 876 occurrences parmi 1 888 583 (hors ponctuation) du corpus, alors que 1 124 677 occurrences de mots de l'antidictionnaire ont été ignorées : le nuage arboré correspond donc à environ 13 % des occurrences du corpus, hors antidictionnaire.

Figure 6 : Chaînes de Formes Partagées des 10 mots les plus fréquents du sous-corpus des scientifiques et du sous-corpus des gens de lettres




La visualisation de la figure 6, inspirée des diagrammes de Sankey[22], aide à détecter des « Chaînes de Formes Partagées » (CFP). Elle montre les deux listes de mots fournies au logiciel, disponible sur http://treecloud.org/cfp/index_fr.php. Les tailles des mots dépendent de leur nombre d'occurrences dans les deux listes fournies[23], normalisées de telle manière que le mot le plus fréquent ait la même taille dans chacune des deux listes, de même pour le mot le moins fréquent affiché dans chacune des deux listes. Si le même mot apparaît du côté gauche et du côté droit, un lien est dessiné entre les deux occurrences[24]. Il est également facile de visualiser les mots présents parmi les plus fréquents dans un sous-corpus sans l'être dans l'autre : ceux qui ne sont associés à aucun lien. Enfin, les intersections de liens permettent de repérer immédiatement les inversions de formes partagées, comme « vie »/« nature », « homme » et « corps », gradation allant

22 Patrick Riehmann, Manfred Hanfler et Bernd Froehlich, « Interactive Sankey Diagrams », *Proceedings of the IEEE Symposium on Information Visualization (INFOVIS 2005)*, 2005, p. 233-240.

23 Normalisées en fonction du nombre d'occurrences le plus faible et le plus élevé.

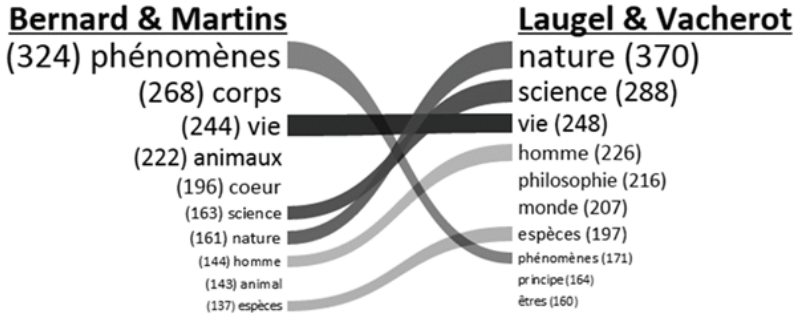
24 Dans l'interface web, le lien est coloré en bleu si le mot est mieux classé dans la seconde liste, rouge s'il est mieux classé dans la première, gris s'il a le même classement dans les deux listes.

du concept général à l'être plus matériel et individuel, qui vont guider l'interprétation contrastive des textes présentée ci-après. Nous limitons ces CFP aux 10 mots les plus fréquents pour des raisons de place, mais il est possible d'en construire sur un nombre plus important de mots.

Paradoxalement, le terme « science » est rangé plus haut chez les gens de lettres que chez les scientifiques. Cela est cohérent avec le fait que les termes les plus fréquents chez les scientifiques concernent essentiellement les objets d'étude, parties du vivant (animaux, corps, espèces, vie, homme) ou de leur environnement (temps, air, eau, nature), alors que les principales préoccupations des seconds portent plutôt sur des concepts généraux qui méritent une réflexion et une discussion alimentée par les découvertes de la science. Les cooccurrents des formes « science » et « sciences » seront traités dans la section suivante.

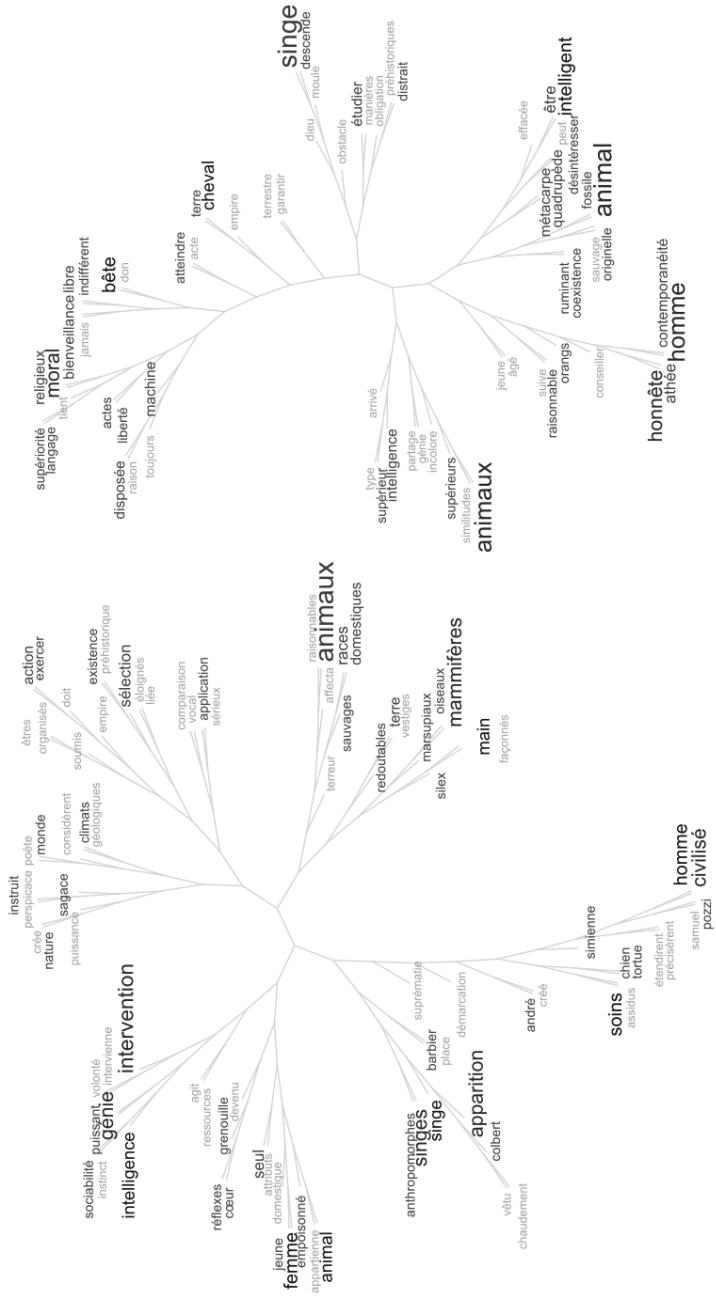
Appliquée à un petit corpus de textes philosophiques et scientifiques également tirés de la *Revue des deux mondes*, afin de consolider nos analyses sur un sous-corpus de test de 20 textes (cinq par auteur), une CFP permet de comparer les deux listes de mots les plus fréquents de deux auteurs scientifiques, Claude Bernard (1813-1878) et Charles Martins (1806-1889), et deux hommes de lettres, Étienne Vacherot (1809-1897) et Auguste Laugel (1830-1914), et de visualiser des ensembles de mots classés dans un ordre inversé dans la première et la seconde liste. Pour trouver de tels ensembles, il faut trouver des liens qui s'intersectent tous, dont les extrémités gauche et droite apparaissent selon un ordre inversé. Il est frappant de constater qu'on peut visualiser en un clin d'œil, par un traducteur de fréquences, les deux angles d'approche d'un même objet, reflétant des méthodologies différentes dans la démarche scientifique et philosophique.

Figure 7 : CFP des 10 mots les plus fréquents d'un sous-corpus de scientifiques et d'un sous-corpus d'hommes de lettres, dans laquelle la plus longue inversion (« phénomènes », « vie », « science », « nature ») apparaît en teintes sombres




Comparons à présent les cooccurrents des mots « homme » et « corps » dans les deux sous-corpus. Dans la figure 8, le vocable des deux sous-corpus autour de l'« homme » indique de nouveau des centres d'intérêt différents : à gauche, les scientifiques discutent les rapports de l'Homme aux autres mammifères d'un point de vue évolutionniste (la partie de droite de l'arbre, ainsi qu'un sous-arbre à gauche) et étudient certaines qualités humaines, réunies dans un même sous-arbre (« génie », « instinct », « intelligence », « sociabilité », « volonté »). À droite, le nuage des cooccurrents du mot homme met en avant les préoccupations des gens de lettres pour « la grande chaîne de la vie », mais en hiérarchisant les rapports de l'Homme à l'animal dans l'échelle des êtres (répétition de « supérieur »/« supériorité », proximité de « animal »/« quadrupède » avec « intelligent », « orangs » avec « raisonnable », « animaux » avec « intelligence »), adossés à des questionnements moraux et religieux (« moral », « religieux », « athée », « dieu »).

Figure 8 : Cooccurrents de « homme » dans les deux sous-corpus (scientifiques à gauche, gens de lettres à droite)



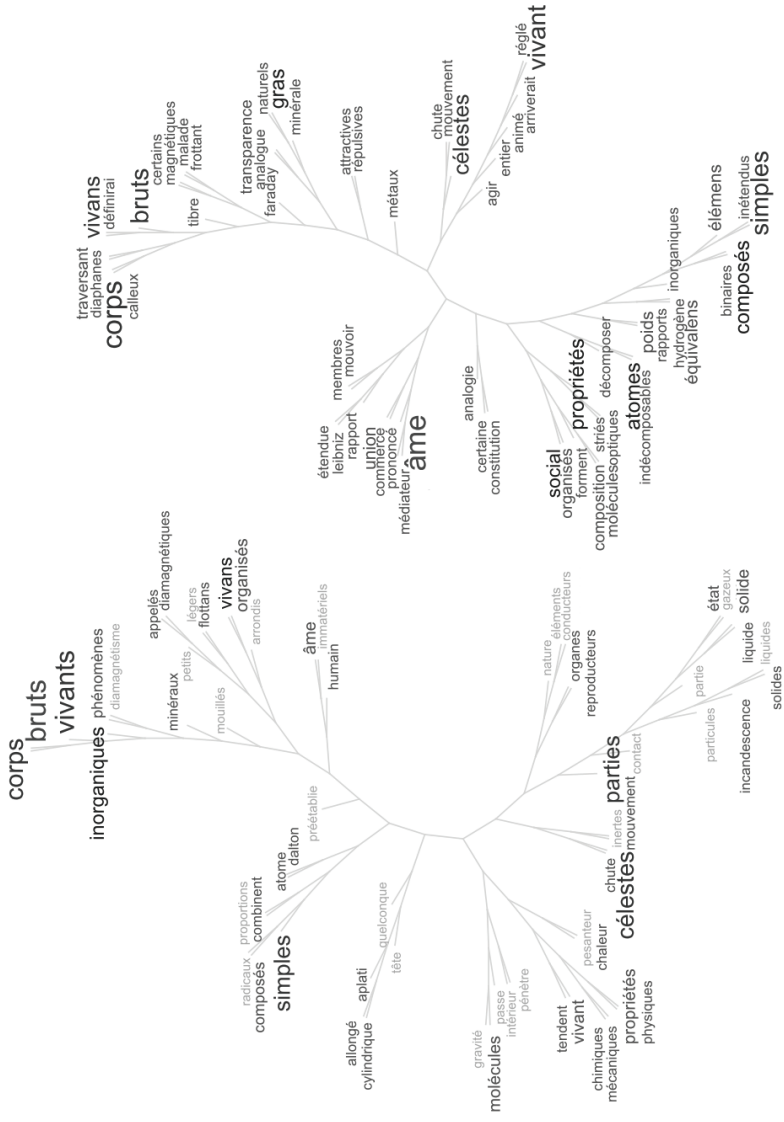
On retrouve presque logiquement ces différences de positionnement dans les nuages des cooccurrents du mot « corps » qui opèrent comme un zoom en figure 9. Le nuage des textes scientifiques fait ressortir un focus sur l'infiniment petit (« petits », « arrondis », « particules », « molécules », théorie atomique de Dalton), les phénomènes physico-chimiques, le mouvement (« chute », « flottans », « pénètre ») et l'état (« solide », « liquide », « gazeux », etc.) des corps. En face, le nuage des gens de lettres semble à première vue partager les mêmes formes, ce qui révèle une sélection de ces aspects. Mais la thématique des rapports entre l'âme et le corps fait son apparition, avec la vie et la matière, d'une part (la présence de « Leibniz » à proximité de « âme » nous met sur la voie). D'autre part, les débats autour de l'organisme et du mécanisme (« libre », « génie », « intelligent »/« intelligence », « langage », « bête », « machine » dans les nuages du mot « homme »), et la science et la religion (nuages arborés des mots « homme » et « corps »), confirment leur place parmi les thématiques chères à certains écrivains du corpus Biolographes comme l'ont amplement démontré Gisèle Séginger[25] et Juliette Azoulai[26]. Des formes partagées par les deux sous-corpus revêtent ainsi des sens différents, comme l'illustre le mot « corps » de l'expression « corps social », qu'Alfred Fouillée évoque dans deux articles, en la mettant en relation avec le corps au sens propre (pour

25 Gisèle Séginger, « Éléments pour une biocritique », *Flaubert*, nº 13, 2015, https://flaubert.revues.org/2439, site consulté le 17 mai 2016; Gisèle Séginger, « Louis Bouilhet et Flaubert. L'invention d'une nouvelle poésie scientifique », dans Muriel Louâpre, Hugues Marchal et Michel Pierssens (dir.), *La poésie scientifique, de la gloire au déclin*, 2014, p. 361-377; Gisèle Séginger, « La réécriture de Cuvier : la création du monde entre savoir et féerie », dans Stéphanie Dord-Crouslé (dir.), « Les dossiers documentaires de Bouvard et Pécuchet » : l'édition numérique du creuset flaubertien. Actes du colloque de Lyon, 7-9 mars 2012, *Flaubert*, nº 13, 2013, http://flaubert.univ-rouen.fr/revue/revue13/documents/Gustave_Flaubert_revue_13_article_Gisele_Seginger.pdf, site consulté le 17 mai 2016; Gisèle Séginger (dir.), Le vivant, Romantisme – La revue du XIXe siècle, vol. 154, nº 4, 2011.

26 Juliette Azoulai, « De la rage métaphysique au calme scientifique : religion et sciences naturelles chez Flaubert », *Flaubert*, nº 13, 2015, https://flaubert.revues.org/2432, article consulté le 17 mai 2016; Juliette Azoulai, *L'âme et le corps chez Flaubert. Une ontologie simple*, Paris, Classiques Garnier, 2014.

reprendre et approfondir une analogie scientifique proposée par Rousseau) ou le mot « mouvement ».

Figure 9 : Cooccurrents de « corps » dans les deux sous-corpus (scientifiques à gauche, gens de lettre à droite)



Cette entrée lexicale « mouvement » est deux fois plus fréquente chez les scientifiques, mais son rapport à la notion de « force » est plus important chez les gens de lettres (indice de cooccurrence avec mouvement 5; 4 chez les scientifiques), où l'idée de force n'est pas seulement consacrée, comme chez les scientifiques, à la matière (parmi les cooccurrents de « mouvement » chez les scientifiques, on a « moléculaire »/« molécules », « vibratoire », « rotation », « masse », « rapide »/« accéléré »). Le philosophe Paul Janet (1823-1899), qui traite de la philosophie leibnizienne fondée sur la notion de force et de force individuelle, n'écrit-il pas :

> Les forces qui composent le corps sont donc des élémens simples, inétendus, des atomes incorporels. Ainsi l'univers est un vaste dynamisme, un savant système de forces individuelles, harmoniquement liées sous le gouvernement d'une force primordiale, dont l'activité absolue laisse subsister en dehors d'elle l'activité propre des créatures et les dirige sans les absorber[27] ?

## 3. L'esthétique des savoirs biologiques

Des formes partagées par les deux sous-corpus revêtent des sens différents, certes, mais comment expliquer la sélection et le transfert de certains savoirs sur d'autres ? Au-delà du repérage des grandes thématiques, il nous intéresse d'offrir des éléments d'analyse portant sur les spécificités des thèmes et formes des savoirs biologiques dans le corpus de réception et de vulgarisation, afin de faire le lien avec leurs variations dans le corpus principal des œuvres littéraires. Ces formes et thèmes sont bien connus des chercheurs littéraires travaillant au sein du projet *Biolographes*; elles comprennent comme nous l'avons évoqué précédemment les éléments de l'infiniment petit et de l'infiniment grand, et par conséquent, les polypes, les infusoires, les gemmules, le plasma, les monères et autres « merveilles » du vivant que l'on trouve dans les thématiques de la création du monde, des origines du vivant ou de la naissance de la vie.

---

27 Paul Janet, « L'idée de force et la philosophie dynamiste », *Revue des deux mondes*, t. 3, 1874, p. 77-107.

Hugues Marchal, qui a étudié la circulation des savoirs scientifiques dans ses travaux sur la poésie scientifique et la vulgarisation au XIX^e^ siècle, souligne que

> chaque genre littéraire impose lui aussi un « modèle du monde » [et] offre « un système complexe de moyens et de manières de prendre possession de la réalité » [...] Mais cette tension résulte dès lors moins d'une contradiction interne que de la structure générique particulière de la vulgarisation, où coexistent deux modèles de composition, dont l'un, scientifique au sens restreint, interdit l'anthropomorphisme, tandis que l'autre, esthétique, méditatif ou pédagogique, le favorise[28] .

Il convoque l'idée de « patron esthétique » propre aux savoirs biologiques : « L'embellissement du savoir n'est pas produit par la seule énonciation. Il résulte d'abord d'un butinage ou raffinage qui ramasse dans le discours scientifique la matière, et parfois le patron esthétique, de tableaux sévèrement limités[29] ».

Nous faisons ainsi l'hypothèse que la sélection des savoirs dans les textes de réception et de vulgarisation procède d'un premier tri de ces derniers sur leur « esthétique première » (ou « patron esthétique » selon Hugues Marchal), donnant ainsi peut-être des pistes quant à la façon dont les savoirs biologiques sont sélectionnés puis transformés, remaniés de nouveau, dans les œuvres littéraires, sans que ces textes de réception et de vulgarisation constituent eux-mêmes de véritables morceaux d'écriture littéraire. Ce sont les modalités (chez les scientifiques et les gens de lettres) de cette esthétique de la science que nous cherchons à identifier dans les visualisations des figures 10 et 11.

---

28 Hugues Marchal, « Le conflit des modèles dans la vulgarisation entomologique : l'exemple de Michelet, Flammarion et Fabre », *Romantisme*, vol. 138, n^o^ 4, 2007, p. 61-74.

29 Hugues Marchal, « L'ambassadeur révoqué : poésie scientifique et popularisation des savoirs au XIX^e^ siècle », *Romantisme*, vol. 144, n^o^ 2, 2009, p. 27.

Figure 10 : Cooccurrents de « sciences » (à gauche) et « science » (à droite) dans le sous-corpus des auteurs scientifiques

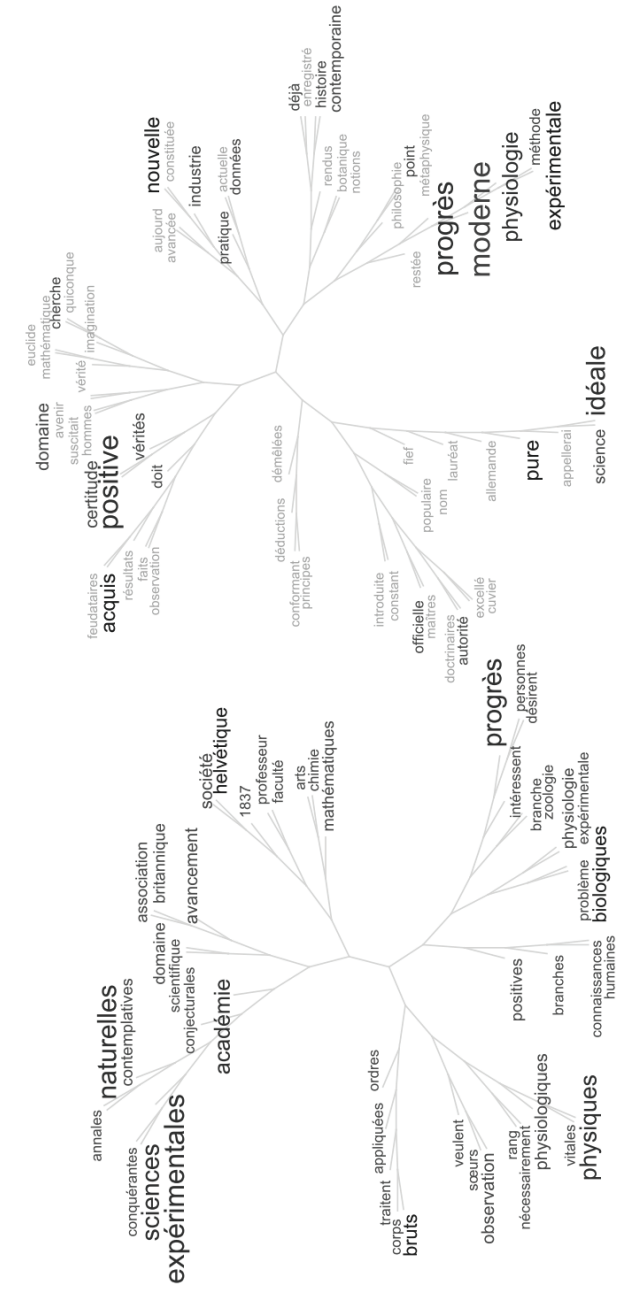
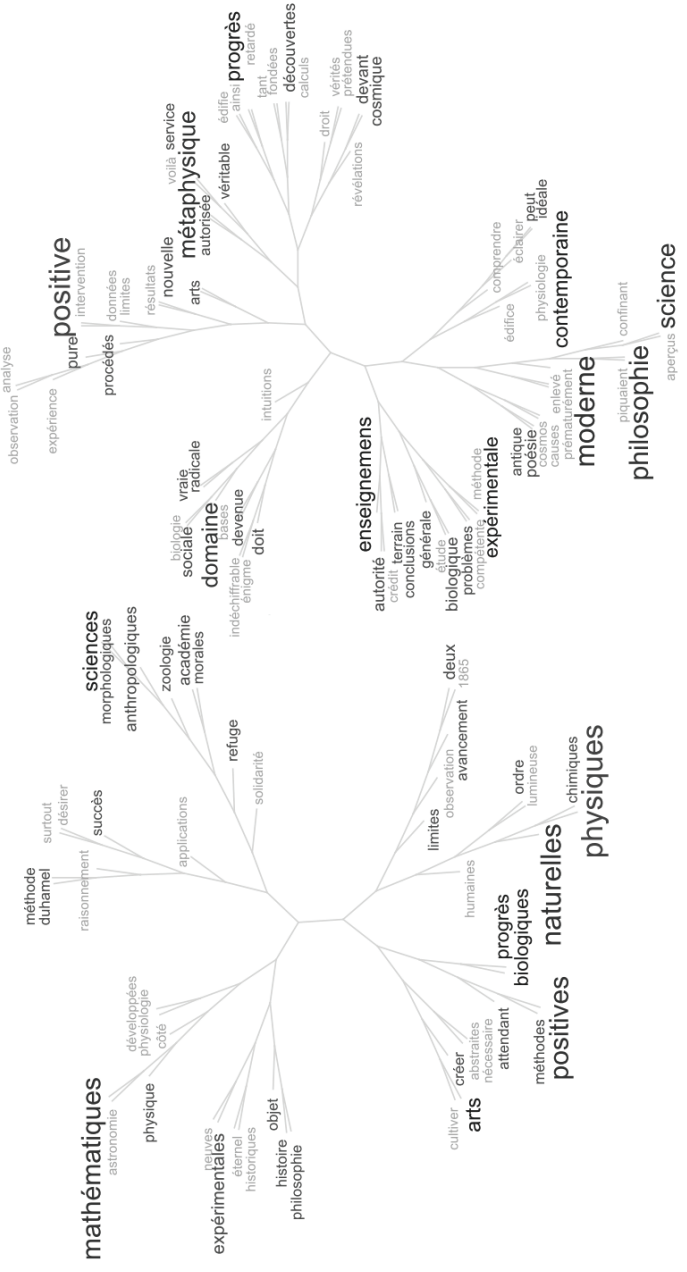


Figure 11 : Cooccurrents de « sciences » (à gauche) et « science » (à droite) dans le sous-corpus des textes écrits par les gens de lettres



Les disciplines scientifiques sont facilement identifiables chez les scientifiques comme chez les gens de lettres, dans le nuage de « sciences » comme de « science » (sciences expérimentales, naturelles, physiologiques, mathématiques; la physique, la chimie, la zoologie, etc.), mais plusieurs sciences humaines font leur apparition dans les nuages des gens de lettres (philosophie, métaphysique, anthropologie, histoire, poésie). Les deux sous-corpus partagent les formes « progrès » et « arts ». Il n'est pas étonnant de voir ce mot apparaître dans les deux sous-corpus compte tenu de l'usage qu'on en fait au XIX$^{e}$ siècle. Les termes d'« art » ou « arts » sont utilisés couramment au sens de « technique », par opposition à la science, conçue comme une pure connaissance indépendante des applications. La liste des collocations de la forme « art » dans le corpus entier renvoie aussi bien aux « arts chimiques, industriels, mécaniques, pratiques », qu'à l'« art militaire, médical, agricole, vétérinaire ». Claude Bernard écrivait au sujet de la médecine : « La médecine n'est pas une science; c'est un art; par conséquent, son application est inséparable de l'artiste. [...] La science est dans la recherche des lois des phénomènes et dans la conception des théories; l'art est dans l'application [...][30] ».

La forme « progrès », elle, est associée dans le corpus des textes scientifiques à des sciences appliquées (voire à l'industrie) et au développement de la méthode expérimentale, alors que, dans le corpus des gens de lettres, elle est plus souvent associée au progrès de l'esprit humain. On trouve d'ailleurs dans le voisinage de progrès parmi les cooccurrents de « science » le terme « révélations », utilisé à plusieurs reprises par Vacherot pour exprimer la fascination devant les découvertes de la science, ou par Caro qui reprend des réflexions de Littré sur la révélation théologique et la révélation scientifique. L'expression « énigme indéchiffrable », dans le voisinage de « biologie » (sous-nuage à droite dans les cooccurrents de « science » chez les gens de lettres), montre davantage encore des perspectives bien différentes chez les scientifiques et les gens de lettres face à la nature. Chez le paléo-

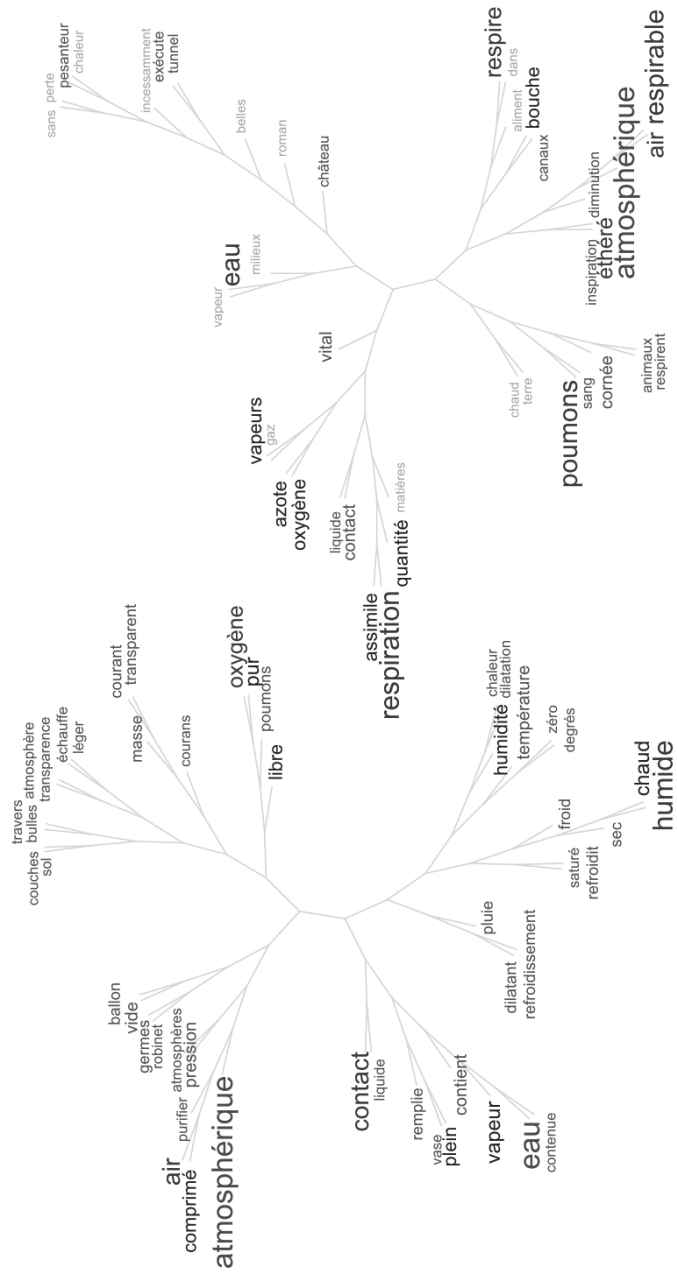
30 Claude Bernard, *Principes de médecine expérimentale*, Paris, Émile Martinet, 1867, p. 175.

botaniste Gaston de Saporta (1823-1895), c'est la théorie de l'évolution qui permet de donner un sens au passé : le passé n'est qu'une « énigme indéchiffrable » sans la théorie de l'évolution. En revanche, la même expression est utilisée chez les philosophes Charles Lévêque (1818-1900) et Étienne Vacherot (1809-1897) (qui signe une sorte d'apologie du vitalisme), en défense de la métaphysique et du finalisme sans lesquels l'évolution demeurerait une « énigme indéchiffrable ».

L'air que nous respirons, ou comme l'écrit George Sand (1804-1876) dans le corpus « Je te vois et je te sens dans tout[31] », voilà l'entrée thématique principale dans les textes des gens de lettres (voir la figure 12) : l'air « vital », celui que « respire[nt] » « bouche » et « poumons ». Ce focus nous rappelle les cooccurrents du mot air dans un échantillon du corpus principal centré sur six œuvres de Jules Michelet et Edgar Quinet, où l'indice de cooccurrence entre les lemmes « air » et « respirer » était également élevé, en particulier dans *La création* où les animaux tirent leur substance alternativement de la mer et de l'« air atmosphérique ». La présence de formes inattendues tels « belles » ou « roman » proviennent de l'usage dans les textes des gens de lettres de nombreuses locutions figuratives utilisant le mot air (comme dans « avoir l'air »). Ces éléments, mis en contraste avec le nuage de gauche représentant les textes scientifiques, où l'air est appréhendé de tous côtés, par ses propriétés, ses effets et sa nature, introduisent (et offrent aux lecteurs) une première forme de sélection sur l'esthétique dans la façon de rapporter et décrire les connaissances scientifiques.

31 George Sand, « Lettres d'un voyageur à propos de botanique », *Revue des deux mondes*, t. 76, 1868, p. 470-496.

Figure 12 : Cooccurrents de « air » dans le sous-corpus des scientifiques (à gauche) et gens de lettres (à droite)

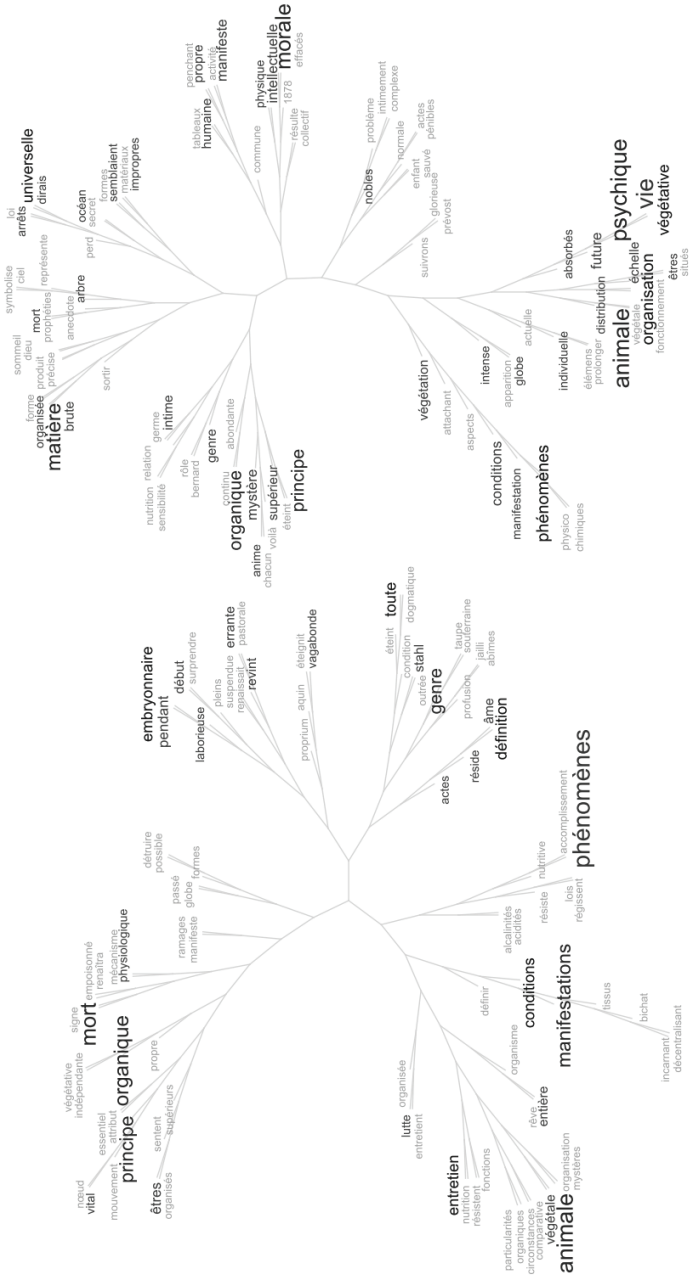


Enfin, les deux derniers nuages de la figure 13 autour du mot « vie » viennent appuyer les analyses des cooccurrents des formes « homme » et « corps » faites précédemment (voir la partie 2). La vie et la mort sont bien sûr présents dans les deux nuages et l'on retrouve dans le nuage des gens de lettres, en haut à droite, les aspects « physico-chimique » (la « matière » « brute », les « éléments »), des « phénomènes » de la vie davantage abordés, dans ce sous-corpus, sous l'angle du questionnement scientifique sur les caractéristiques de la vie (« phénomènes », « conditions », « manifestations », d'une part, « animale » et « végétale », d'autre part). L'apparition de Georg Ernst Stahl (1659-1734) dans le nuage des scientifiques (en haut à droite) tient d'ailleurs du débat autour de la « définition » de la vie, très différente chez les vitalistes. Claude Bernard sur ce point décrivait un Stahl très dogmatique :

> L'animisme a été l'expression outrée de la spiritualité de la vie; Stahl fut le partisan déterminé et le plus dogmatique de ces idées perpétuées depuis Aristote. [...] Stahl comprit tout autrement la nature des phénomènes de la vie et les rapports de l'âme et du corps. Dans les actes vitaux, il rejette toutes les explications qui leur seraient communes avec les phénomènes mécaniques, physiques et chimiques de la matière brute[32].

[32] Claude Bernard, « Définition de la vie. Les théories anciennes et la science moderne », *Revue des deux mondes*, t. 9, 1875, p. 326-349.

Figure 13 : Cooccurrents de « vie » dans le sous-corpus des scientifiques (à gauche) et des gens de lettres (à droite)



Le nuage des gens de lettres (à droite) se construit lui aussi en écho au nuage des cooccurrents des mots « homme » et « corps » dans le même sous-corpus, qu'il s'agisse de l'évocation de la grande chaîne de la vie (autour d'« animale » en bas, « organisation », « êtres », « échelle ») ou du corps lui-même (en haut autour de « matière »). Mais alors que la thématique de la vie psychique n'est visible qu'à travers les termes « âme », « rêve » et « végétative » dans l'arbre de gauche, elle l'est beaucoup plus dans celui de droite (« morale », « intellectuelle », « psychique », « végétative »). On y trouve du vocabulaire qui nous éloigne d'une approche purement scientifique : « secret », « mystère » ou « prophéties » sont autant de termes liés à l'esthétique de la religion, qui apparaît d'ailleurs explicitement à gauche de l'arbre avec le mot « dieu ». Un retour au texte permet de constater que le terme est utilisé à proximité de « vie » dans des passages qui décrivent des démarches contraires à la démarche scientifique. Ces passages sont particulièrement riches en métaphores, centrées autour du dieu grec Pan chez George Sand, de l'« image de la vie » que constituent les « fleurs lancées par l'arc du dieu de l'amour[33] » chez Dora d'Istria, ou encore de l'image du « dieu capricieux et de Protée menteur, échappant à toute prise[34] » que revêt la vie pour les médecins superstitieux selon Paul Janet.

## 4. Conclusion

Dans les deux parties qui précèdent, nous avons montré la complémentarité des outils textométriques, comme TXM, et visuels, comme TreeCloud et le nouveau traducteur de fréquences colorées introduit dans cet article (voir les figures 6 et 7). Ces différentes approches permettent un travail sur les fréquences et spécificités d'un ensemble de textes plus ou moins important, où le retour au texte vient toujours enrichir les interprétations. Les nuages de mots arborés produits par TreeCloud ont affiné l'analyse contrastive en fournissant des visualisations de mots-clés et leurs

33 Dora d'Istria, « Le surnaturel dans le monde végétal », *Revue des deux mondes*, t. 32, 1874, p. 481-508.

34 Paul Janet, « La méthode expérimentale et la physiologie à propos des travaux récens de M. Claude Bernard », *Revue des deux mondes*, 1866, p. 908-936.

cooccurrents jouant les mêmes rôles ou des rôles différents dans les deux sous-corpus.

D'une façon générale, certains nuages ont mis en valeur les modalités possibles de la sélection des savoirs, en particulier, l'idée que plus la description scientifique d'un objet est ténue (description fine de la matière, de l'air, des corps), plus son « empreinte esthétique » est grande, plus cet objet a de chance d'être sélectionné et remanié dans les textes de réception et de vulgarisation. Il nous faut encore documenter plus finement ce rapport entre description scientifique fine et sélection des objets dans les textes de réception et les textes littéraires par un travail sur les fréquences de certains vocables : structures argumentales des adjectifs et des verbes supports en particulier[35], études des formes composées et segments répétés. Ce travail trouve dès lors un prolongement naturel dans la recherche de moyens permettant le repérage de constructions stylistiques et l'extraction de métaphores dans les textes de réception et de vulgarisation pour voir comment se construit la poétique de la science (ou une première forme d'esthétique) dans ces écrits (avec l'utilisation d'autres logiciels tel Unitex, Lexico 5 ou TextObserver).

35 Maurice Gross, « Les bases empiriques de la notion de prédicat sémantique », *Langages*, nº 63, 1981, p. 7-52; Maurice Gross, « Une grammaire locale de l'expression des sentiments », *Langue française*, nº 105, 1995, p. 70–87.

## Bibliographie

Amstutz, Delphine et Philippe Gambette, « Utilisation de la visualisation en nuage arboré pour l'analyse littéraire », dans Sergio Bolasco, Isabella Chiari, Luca Giuliano (dir.), *Statistical Analysis of Textual Data, Proceedings of the 10th International Conference on Statistical Analysis of Textual Data (JADT 2010)*, Edizioni Universitarie di Lettere Economia Diritto, 2010, p. 227-238.

Azoulai, Juliette, « De la rage métaphysique au calme scientifique : religion et sciences naturelles chez Flaubert », *Flaubert*, nº 13, 2015.

Azoulai, Juliette, *L'âme et le corps chez Flaubert. Une ontologie simple*, Paris, Classiques Garnier, 2014.

Bernard, Claude, « Définition de la vie. Les théories anciennes et la science moderne », *Revue des deux mondes*, t. 9, 1875, p. 326-349.

Bernard, Claude, *Principes de médecine expérimentale*, Paris, Émile Martinet, 1867.

Dufour, Philippe, « La feuille bulozienne », *Flaubert*, nº 9, 2013, http://flaubert.revues.org/2024, article consulté le 17 mai 2016.

Gambette, Philippe, *User Manual for TreeCloud*, 2010, http://www.treecloud.org/DOWNLOADS/ManualTreecloud.pdf, site consulté le 25 mars 2016.

Gambette, Philippe et Jean Véronis, « Visualising a Text with a Tree Cloud », dans Hermann Locarek-Junge et Claus Weihs (dir.), *Studies in Classification, Data Analysis, and Knowledge Organization, Proceedings of the International Federation of Classification Societies 2009 Conference (IFCS'09)*, nº 40, 2010, p. 561-570.

Gambette, Philippe, Nuria Gala et Alexis Nasr, « Longueur de branches et arbres de mots », *Corpus*, nº 11, 2012, p. 129-146.

Gross, Maurice, « Les bases empiriques de la notion de prédicat sémantique », *Langages*, nº 63, 1981, p. 7–52.

Gross, Maurice, « Une grammaire locale de l'expression des sentiments », *Langue française*, nº 105, 1995, p. 70–87.

Heiden, Serge, Jean-Philippe Magué et Bénédicte Pincemin, « TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement », dans Sergio Bolasco, Isabella Chiari, Luca Giuliano (dir.), *Statistical Analysis of Textual Data, Proceedings of the 10th International Conference on Statistical Analysis of Textual Data (JADT 2010)*, Edizioni Universitarie di Lettere Economia Diritto, 2010, p. 1021-1032.

Istria, Dora d', « Le surnaturel dans le monde végétal », *Revue des deux mondes*, t. 32, 1874, p. 481-508.

Janet, Paul, « L'idée de force et la philosophie dynamiste », *Revue des deux mondes*, t. 3, 1874, p. 77-107.

Janet, Paul, « La méthode expérimentale et la physiologie à propos des travaux récens de M. Claude Bernard », *Revue des deux mondes*, 1866, p. 908-936.

Lafon, Pierre, « Sur la variabilité de la fréquence des formes dans un corpus », *Mots*, vol. 1, n° 1, p. 127-165, 1980.

Lechevrel, Nadège, « Fouille de données textuelles et recherche documentaire automatiques pour l'histoire des théories linguistiques », dans Pascal Charbonnat, Mahé Ben Hamed, Guillaume Lecointre (dir.), *Apparenter la pensée ? Vers une phylogénie des concepts savants*, Matériologiques, 2014, p. 219-243.

Lechevrel, Nadège, *Réception et vulgarisation des savoirs biologiques dans le corpus Biolographes*, 2015, *billet de blog sur le carnet hypotheses.org du projet Biolographes*, http://biolog.hypotheses.org/1276, site consulté le 25 mars 2016.

Marchal, Hugues, « L'ambassadeur révoqué : poésie scientifique et popularisation des savoirs au XIX^e siècle », *Romantisme*, vol. 144, n° 2, 2009, p. 25-37.

Marchal, Hugues, « Le conflit des modèles dans la vulgarisation entomologique : l'exemple de Michelet, Flammarion et Fabre », *Romantisme*, vol. 138, n° 4, 2007, p. 61-74.

Riehmann, Patrick, Manfred Hanfler et Bernd Froehlich, « Interactive Sankey Diagrams », *Proceedings of the IEEE Symposium on Information Visualization (INFOVIS 2005)*, 2005, p. 233-240.

Sand, George, « Lettres d'un voyageur à propos de botanique », *Revue des deux mondes*, t. 76, 1868, p. 470-496.

Schmid, Helmut, « Probabilistic Part-of-Speech Tagging Using Decision Trees », *Proceedings of International Conference on New Methods in Language Processing*, 1994, p. 44-49.

Séginger, Gisèle (dir), *Biolographes*, projet ANR-13-FRAL-0013, 2014-2016, http://biolog.hypotheses.org/.

Séginger, Gisèle, « Éléments pour une biocritique », *Flaubert*, n° 13, 2015, https://flaubert.revues.org/2439, article consulté le 17 mai 2016.

Séginger, Gisèle, « Louis Bouilhet et Flaubert. L'invention d'une nouvelle poésie scientifique », dans Muriel Louâpre, Hugues Marchal et Michel Pierssens (dir.), *La poésie scientifique, de la gloire au déclin*, 2014, p. 361-377.

Séginger, Gisèle, « La réécriture de Cuvier : la création du monde entre savoir et féerie », dans Stéphanie Dord-Crouslé (dir.), « Les dossiers documentaires de Bouvard et Pécuchet » : l'édition numérique du

creuset flaubertien. Actes du colloque de Lyon, 7-9 mars 2012, *Flaubert*, nº 13, 2013, http://flaubert.univ-rouen.fr/revue/revue13/documents/Gustave_Flaubert_revue_13_article_Gisele_Seginger.pdf, site consulté le 17 mai 2016.

Séginger, Gisèle (dir), *Le vivant, Romantisme - La revue du XIXe siècle*, vol. 154, nº 4, 2011